Genet. Res., Camb. (1975), **25**, *pp.* 145–153
With 4 text-figures
Printed in Great Britain

Extensions to the model of an infinite number of selectively neutral alleles in a finite population

By P. J. AVERY

Department of Biomathematics, University of Oxford, Pusey St, Oxford OX1 2JZ

(*Received* 17 *December* 1974)

SUMMARY

The model, suggested by Ohta & Kimura, of an infinite number of selectively neutral alleles, in which alleles can only mutate to neighbouring alleles, has been extended to include a migratory pattern. The stepping-stone and island models of migration have been considered. In the latter case, it has been found that as the number of colonies becomes large, the expected number of alleles, n_e, becomes approximately equal to

$$[(1+4N_em)(1+4N_em+8N_ev)]^{\frac{1}{2}},$$

where m is the migration rate, v is the mutation rate and N_e is the effective size of each population. This solution tends to that of Ohta & Kimura as $m \to 0$ but one can see that for any appreciable value of N_em, a large increase in n_e is obtained. In order to check the validity of iterative results, models with a finite number of alleles have been considered, and their solutions have been found to converge quickly to those of the infinite case. The results exemplify the great power of migratory structure and neutral alleles to maintain a large amount of heterozygosity. Double step mutation and the finite time solution have also been considered.

1. MODEL AND CONSTRUCTION OF EQUATIONS

Let us assume that the entire sequence of allelic states are expressed by integers, i.e. $(\ldots, A_{-1}, A_0, A_1, \ldots)$ and that an allele can only mutate one step in each direction. Consider a diploid population of N colonies each with effective population size, N_e. Let v be the mutation rate per locus per generation and let m be the migration rate away from a colony per locus per generation. Mutation is assumed to be equal in each direction. Let x_{ij} be the frequency of the ith allele in the jth colony and let

$$X_{ik} = E[\sum_l x_{l,j}x_{l+i,j+k}]$$

for all j, where $i = 1, 2, \ldots$, and $k = 1, 2, \ldots, \frac{1}{2}N$ (N even) or $\frac{1}{2}(N-1)$ (N odd).

The basic equation which we shall be using is Kolmogorov's backward equation; this particular form having been derived by Ohta & Kimura (1971) and later expressed for this particular problem in Ohta & Kimura (1973), i.e.

$$\frac{\mathrm{d}}{\mathrm{d}t}[E(f)] = E[L(f)], \tag{1.1}$$

where L is a differential operator and

$$L = \frac{1}{2}\sum_{i,j} V_{\delta x_{ij}} \frac{\partial}{\partial x_{ij}^2} + \sum_{\substack{i>k \\ j>l}} W_{\delta x_{ij}\delta x_{kl}} \frac{\partial^2}{\partial x_{ij}\partial x_{kl}} + \sum_{i,j} M_{\delta x_{ij}} \frac{\partial}{\partial x_{ij}}.$$

To set up the equations, one puts f equal in turn to x_{ij}^2, $x_{ij}x_{i+w,j}$, $x_{ij}x_{i+w,j+z}$, and then sums over all values of i and j for each value of w and z. We are interested in the equilibrium situation which is obtained by putting $\mathrm{d}X_{i,j}/\mathrm{d}t = 0$ for all i, j. The solution must be such that $X_{i,j} \to 0$ as $i \to \infty$ in the infinite allele case.

2. DOUBLE STEP MUTATION

Wehrahn (1974) has suggested that double charge changes happen, but less frequently than single charge changes. It seems reasonable to assume that this double step mutation rate will be a proportion of the single step mutation rate (Kv, say). In this model, only one colony is considered and thus the second suffix of x_{ij} and X_{ij} is not used.

Now $$M_{\delta x_i} = \tfrac{1}{2}v(x_{i-1}+x_{i+1}) + \tfrac{1}{2}Kv(x_{i-2}+x_{i+2}) - v(1+K)x_i,$$

$$V_{\delta x_i} = \frac{x_i(1-x_i)}{2N_e}; \quad W_{\delta x_i \delta x_j} = \frac{-x_i x_j}{2N_e} \quad (i \neq j).$$

(1.1) thus gives

$$1 + 2\beta X_1 + 2\beta K X_2 - (1 + 2\beta(1+K))X_0 = 0, \tag{2.1}$$

$$\beta X_0 - (1 + 2\beta + \beta K)X_1 + \beta X_2 + \beta K X_3 = 0, \tag{2.2}$$

$$K\beta X_{k-2} + \beta X_{k-1} - (1 + 2\beta(1+K))X_k + \beta X_{k+1} + \beta K X_{k+2} = 0, \tag{2.3}$$

for $k \geqslant 2$, where $\beta = 2N_e v$.

On substituting $X_k = X_0\lambda^k$ in (2.3), one obtains

$$\beta K\lambda^4 + \beta\lambda^3 - (1 + 2\beta(1+K))\lambda^2 + \beta\lambda + \beta K = 0. \tag{2.4}$$

This is solved by substituting $\lambda + \lambda^{-1} = z$ and then ignoring the roots of λ, which are greater than one, as $\lambda^k \to 0$ as $k \to \infty$. Thus

$$\lambda_1 = \frac{-1 + \sqrt{\gamma} - \sqrt{(2 + 4K/\beta + 8K - 2\sqrt{\gamma})}}{4K}$$

and

$$\lambda_2 = \frac{-1 - \sqrt{\gamma} + \sqrt{(2 + 4K/\beta + 8K + 2\sqrt{\gamma})}}{4K},$$

where

$$\gamma = 1 + 4K(1 + 2\beta + 4\beta K)/\beta.$$

Then $X_k = {}_1X_0\lambda_1^k + {}_2X_0\lambda_2^k$, and ${}_1X_0$ and ${}_2X_0$ are found by substituting in (2.1) and (2.2).

$$\text{Expected heterozygosity} = H = 1 - {}_1X_0 - {}_2X_0$$

and

$$\text{expected number of neutral alleles} = n_e = \frac{1}{{}_1X_0 + {}_2X_0}.$$

The same method can be applied if the second step mutation rate is independent of the first. One finds that a relatively small value of K still produces a significant change in H and n_e. As $K \to 0$, $n_e \to \sqrt{(1 + 8N_e v)}$ as given in Ohta & Kimura (1973).

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

3. FINITE NUMBER OF ALLELES

Later on in this paper, finite allele approximations have been used. In order to justify their validity, the solution of the finite model has been found. A single colony model, in which there is only single step mutation, has been considered. There are assumed to be K possible alleles arranged on a circle, i.e. x_1 mutates to x_2 and x_K only. (1.1) thus produces

$$(1+2\beta)X_0 - 2\beta X_1 = 1, \tag{3.1}$$

$$\beta X_{j+1} - (1+2\beta)X_j + \beta X_{j-1} = 0 \quad (1 \leqslant j < l), \tag{3.2}$$

$$2\beta X_{l-1} - (1+2\beta)X_l = 0, \quad \text{where } l = K/2, \quad -K \text{ even}, \tag{3.3}$$

$$\beta X_{l-1} - (1+\beta)X_l = 0, \quad \text{where } l = (K-1)/2, \quad -K \text{ odd} \quad (\beta = 2N_e v). \tag{3.4}$$

On putting $X_j = X_0\lambda^j$ in (3.2) one obtains

$$\lambda = \frac{1+2\beta \pm \sqrt{(1+4\beta)}}{2\beta}$$

and on substituting $X_j = {}_1X_0\lambda_1^j + {}_2X_0\lambda_2^j$ in (3.1), and (3.3) or (3.4), one obtains

$$X_j = \frac{1}{\sqrt{(1+4\beta)}} \frac{\lambda^j + \lambda^{K-j}}{1-\lambda^K},$$

where

$$\lambda = \frac{1+2\beta-\sqrt{(1+4\beta)}}{2\beta},$$

and thus effective number of alleles $= n_e = \dfrac{1-\lambda^K}{1+\lambda^K}\sqrt{(1+4\beta)}$.

This tends quickly to $(1+4\beta)^{\frac{1}{2}}$ and agrees to 7 d.p. by $K = 18$. For $K = 2$, $n_e = (1+4\beta/1+2\beta)$, and for $K = 3$, $n_e = (1+3\beta)/(1+\beta)$. This is exemplified by Fig. 1, where $H = 1 - X_0$.

4. ISLAND MODEL OF MIGRATION

The island model was examined recently by Latter (1973), who considered it to be the best model for human populations. The model is that a proportion, m, of each colony migrates away and is then equally shared between each of the other colonies. N colonies, each of effective size N_e, are considered. m and v are considered to be of order (N_e^{-1}) so that quadratic terms in them can be ignored.

Thus

$$M_{\delta x_{ij}} = \frac{v}{2}(x_{i+1,j} + x_{i-1,j} - 2x_{ij}) + \frac{m}{N-1}\left(\sum_{k \neq j} x_{ik} - (N-1)x_{ij}\right),$$

$$V_{\delta x_{ij}} = \frac{x_{ij}(1-x_{ij})}{2N_e};$$

$$W_{\delta x_{ij} x_{kj}} = \frac{-x_{ij}x_{kj}}{2N_e} \quad (i \neq k);$$

$$W_{\delta x_{ij}\delta x_{kl}} = 0 \quad (j \neq l) \quad \text{for all } i \text{ and } k.$$

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

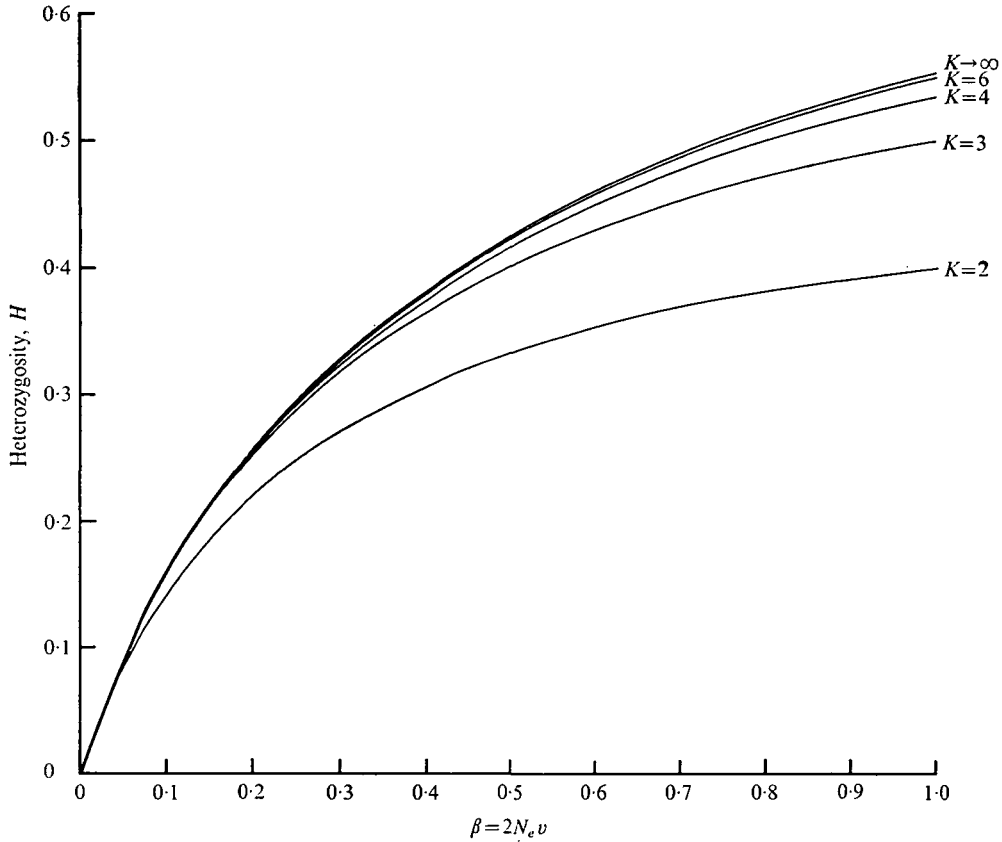

Fig. 1. The steady-state heterozygosity is plotted against β (equals colony size times twice the mutation rate) for the single colony case with a finite number, K, of neutral alleles.

In this model, each colony migrates directly with each other, thus

$$X_{ij} \equiv X_{i,1} \quad \text{where} \quad j \geqslant 1.$$

Thus one obtains by substituting in (1.1) and simplifying

$$1 + 2\beta X_{10} - (1 + M + 2\beta) X_{00} + M X_{01} = 0, \tag{4.1}$$

$$\beta X_{l-1,0} + \beta X_{l+1,0} - (1 + M + 2\beta) X_{l0} + M X_{l1} = 0 \quad (l \geqslant 1), \tag{4.2}$$

$$2\beta X_{11} - \left(M + 2\beta - \frac{M(N-2)}{N-1}\right) X_{01} + \frac{M}{N-1} X_{00} = 0, \tag{4.3}$$

$$\beta X_{l-1,1} + \beta X_{l+1,1} - \left(M + 2\beta - \frac{M(N-2)}{N-1}\right) X_{l1} + \frac{M}{N-1} X_{l0} = 0 \quad (l \geqslant 1), \tag{4.4}$$

where $M = 4N_e m$. On putting $X_{il} = X_l \lambda^i$ for $l = 0, 1$ in (4.2) and (4.4), one obtains

$$\begin{pmatrix} f(\lambda) - \lambda & M\lambda \\ \dfrac{M\lambda}{N-1} & f(\lambda) + \dfrac{M(N-2)\lambda}{N-1} \end{pmatrix} \begin{pmatrix} X_0 \\ X_1 \end{pmatrix} = \mathbf{MAX} = \mathbf{0},$$

where $f(\lambda) = \beta\lambda^2 - \lambda(M + 2\beta) + \beta$ and $\mathbf{0}' = (0, 0)$.

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

For a non-zero solution of $\mathbf{X}$, $|\mathbf{MA}| = 0$. This leads to four roots for λ but the condition that $\lambda^k \to 0$ as $k \to \infty$, implies that two roots must be rejected. Thus

$$\lambda_i = \frac{1}{2\beta}(M + 2\beta + y_i - \sqrt{[(M + 2\beta + y_i)^2 - 4\beta^2]}) \quad (i = 1, 2),$$

where $$y_i = \frac{1}{2}\left(1 - \frac{M(N-2)}{N-1} \pm \sqrt{\left[\left(1 + \frac{M(N-2)}{N-1}\right)^2 + \frac{4M^2}{N-1}\right]}\right) = \frac{f(\lambda_i)}{\lambda_i}.$$

If $\beta \to 0$, $\lambda_i \to 0$ for $i = 1, 2$ and all values of M and N. If

$$N \to \infty \ (\beta \neq 0), \quad \lambda_1 \to \frac{1}{2\beta}(1 + M + 2\beta - \sqrt{[(1 + M + 2\beta)^2 - 4\beta^2]})$$

and $\lambda_2 \to 1$.

Thus, from the matrix, ${}_iX_1 = (1 - y_i)_i X_0/M$ for $i = 1, 2$, where $X_0 = {}_1X_0 + {}_2X_0$. By substituting the above in the initial equations, a pair of linear equations in ${}_1X_0$ and ${}_2X_0$ are found and thus X_0, $n_e\ (= 1/X_0)$ and the expected heterozygosity, $H(= 1 - X_0)$ can be found. As $N \to \infty$ $(\beta \neq 0)$, $y_1 \to 1$ and $y_2 \to -M$ which implies that, for large N, ${}_1X_1 \simeq 0$ and ${}_2X_1 \simeq [(1 + M)/M]\, X_0$. Substitution in (4.1) and (4.3) gives ${}_1X_0 = [(1 + M)(1 + M + 4\beta)]^{-\frac{1}{2}}$ and ${}_2X_0 = 0$. Thus for large N,

$$n_e \simeq \sqrt{[(1 + M)(1 + M + 4\beta)]} \quad (\beta \neq 0) \quad \text{and} \quad n_e = 1 \quad (\beta = 0).$$

The results are given in Fig. 2.

The above limiting conclusions were confirmed by taking very large values for N in the set of four linear equations.

The finite allele approximation was used to form a set of matrix equations. The results obtained from this second approach agreed exactly with the results from above. One sees from the graph that, for fixed $N_e m$, one obtains a large spread of values as N varies. The effect on H of increasing N is large when N is small, but decreases as N becomes large, so that the graph for H tends to a limiting curve. The limiting curves are plotted for $4N_e m = 1{\cdot}0, 2{\cdot}0, 3{\cdot}0, 4{\cdot}0$ showing that $N_e m$ also has a large effect on H.

5. STEPPING-STONE MODEL

In this model, N colonies are arranged on a circle and a proportion m of each colony migrates at each generation, being equally divided between the two adjacent colonies (Kimura & Weiss, 1964). Thus

$$M_{\delta x_{ij}} = \tfrac{1}{2}v(x_{i+1,j} + x_{i-1,j} - 2x_{ij}) + \tfrac{1}{2}m(x_{i,j+1} + x_{i,j-1} - 2x_{ij});$$

$$V_{\delta x_{ij}} = \frac{x_{ij}}{2N_e}(1 - x_{ij});$$

$$W_{\delta x_{ij}\delta x_{kj}} = \frac{-x_{ij}x_{kj}}{2N_e} \quad (i \neq k);$$

$$W_{\delta x_{ij}\delta x_{kl}} = 0 \quad \text{for all } i, k \quad (j \neq l).$$

On substituting in (1.1), one obtains single equations and a set of general equations, which on substitution of $X_{ik} = X_k \lambda^i$ $(k = 0, 1, \ldots, j$ where $j = N/2$ (N even) or $(N-1)/2$ (N odd)) leads to

$$\mathbf{MAX} = 0,$$

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

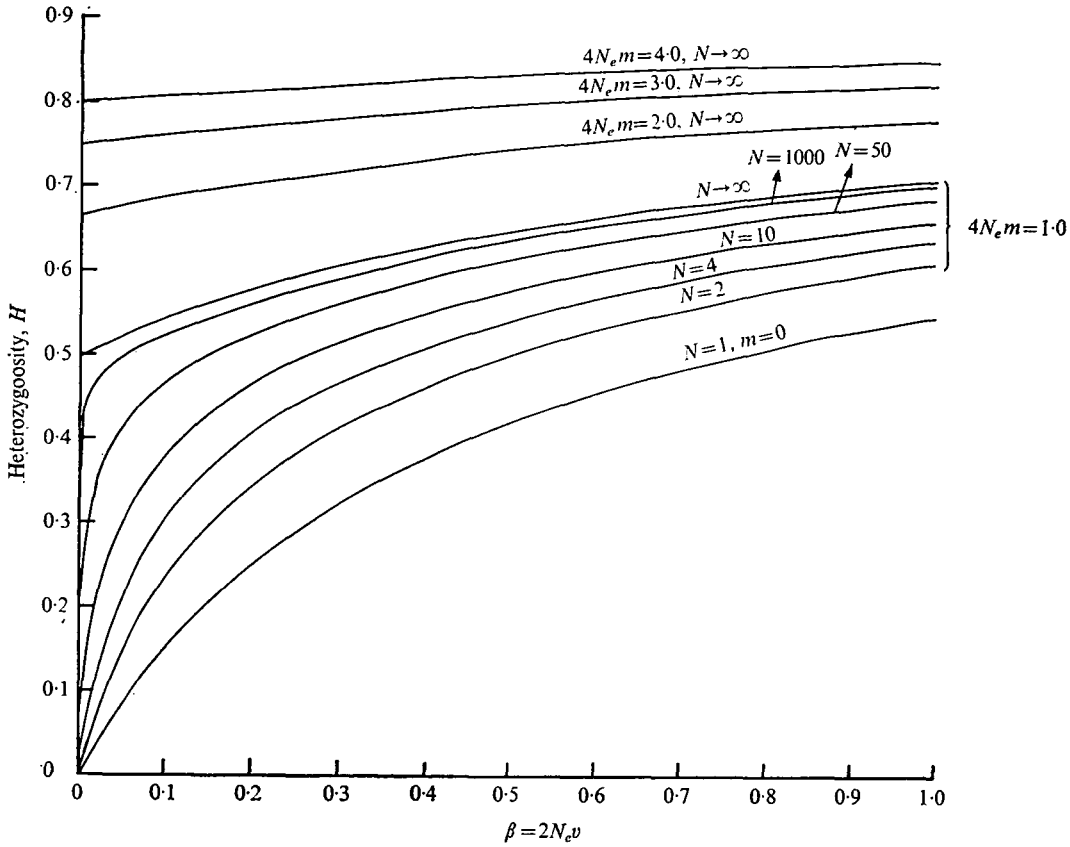

Fig. 2. For the island model of migration, the steady-state heterozygosity in each colony is plotted against the mutation factor, β, for different values of the migration factor, M (four times colony size times migration rate) and for different numbers of colonies, N. The limiting solutions are also shown.

where $\mathbf{X}' = (X_0, X_1, \ldots, X_j)$ and $\mathbf{MA} = f(\lambda)\mathbf{I} + \frac{1}{2}M\lambda\mathbf{R}_1 + \frac{1}{2}M\lambda\mathbf{R}_2 + \mathbf{C}$, where $\mathbf{R}_1$ is a $(j+1)\times(j+1)$ matrix with 1's in super-diagonal and zeros elsewhere, $\mathbf{R}_2$ is a $(j+1)\times(j+1)$ matrix with 1's in sub-diagonal and zeros elsewhere, $f(\lambda) = \beta(\lambda-1)^2 - M\lambda$ and $\mathbf{C}$ is zero everywhere except that $C[1,1] = -\lambda$, $C[1,2] = \frac{1}{2}M\lambda$, $C[j+1,j] = \frac{1}{2}M\lambda$ if N even, $C[j,j] = \frac{1}{2}M\lambda$ if N is odd. For non zero solution of $\mathbf{X}$, $|\mathbf{MA}| = 0$. This leads to $(j+1)$ acceptable roots for λ. Then one puts

$$X_{s,k} = \sum_{i=0}^{j} {}_iX_k\lambda_i^s,$$

and by using j of the original equations, one finds

$$n_e = \left(\sum_{i=0}^{j} {}_iX_0\right)^{-1}.$$

λ_i and ${}_iX_k$ are best found by using methods for evaluation of eigenvalues and eigenvectors. For $N = 2$ and 3, the results are obviously the same as in the island

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

model. However, as N increases, H increases much more slowly than in the island model and converges quickly to a limiting curve, particularly for $\beta > 0{\cdot}1$. Thus this migratory structure has a much smaller effect on heterozygosity for the same value of $4N_em$ as does the island model as can be seen by Fig. 3.

As before, finite allele approximation was used as a check on the above theory and it agreed exactly.

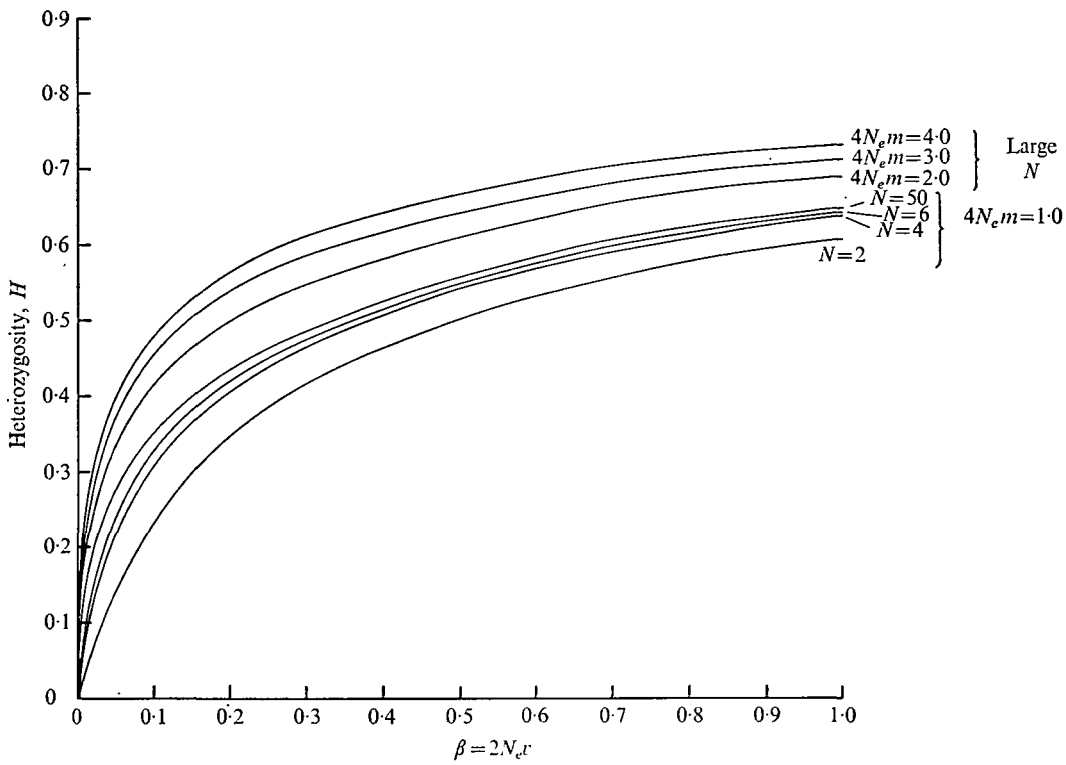

Fig. 3. For the stepping-stone model of migration, the steady-state heterozygosity in each colony is plotted against the mutation factor, β, for different values of the migration factor, M and for different numbers of colonies, N.

6. TRANSIENT SOLUTION FOR THE MODEL OF A FINITE NUMBER OF ALLELES

In the previous sections of this paper, the steady-state solution has always been examined. It seems appropriate that, in conclusion, the transient solution should be examined to see how quickly the steady state solution is reached.

The simplest case is that of a single colony with one step mutation and a finite number of alleles. From sections (1) and (3), one can see that the equations will be:

$$\mathbf{MAX} = \mathbf{V} + \frac{d\mathbf{X}}{dT},$$

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

where $\mathbf{X}' = (X_0, X_1, \ldots, X_{l-1}, X_l)$,

$\mathbf{V}' = (-1, 0, 0, \ldots, 0, 0)$,

T is time measured in $2N_e$ generations, and

$$\mathbf{MA} = -(1+2\beta)\mathbf{I} + \beta\mathbf{R}_1 + \beta\mathbf{R}_2 + \mathbf{C}.$$

$\mathbf{I}$, $\mathbf{R}_1$, $\mathbf{R}_2$ and $\mathbf{C}$ are all $(l+1) \times (l+1)$ matrices where $\mathbf{I}$ is the identity matrix, $\mathbf{R}_1$ and $\mathbf{R}_2$ are as defined in section 5 and $\mathbf{C}$ is a zero matrix except that

$$C[1,2] = \beta,\ C[l+1, l] = \beta \text{ if } K \text{ is even},\ C[l, l] = \beta \text{ if } K \text{ is odd}.$$

If one substitutes $\mathbf{Y} = \mathbf{X} - \mathbf{S}$, where $\mathbf{S}$ is the steady state solution, one obtains $\mathbf{MA}\,\mathbf{Y} = \mathrm{d}\mathbf{Y}/\mathrm{d}T$ whose solution is

$$\mathbf{Y} = e^{\mathbf{MA}t} = \sum_{i=0}^{l} \boldsymbol{\gamma}_i \boldsymbol{\delta}_i e^{\lambda_i t}\mathbf{E},$$

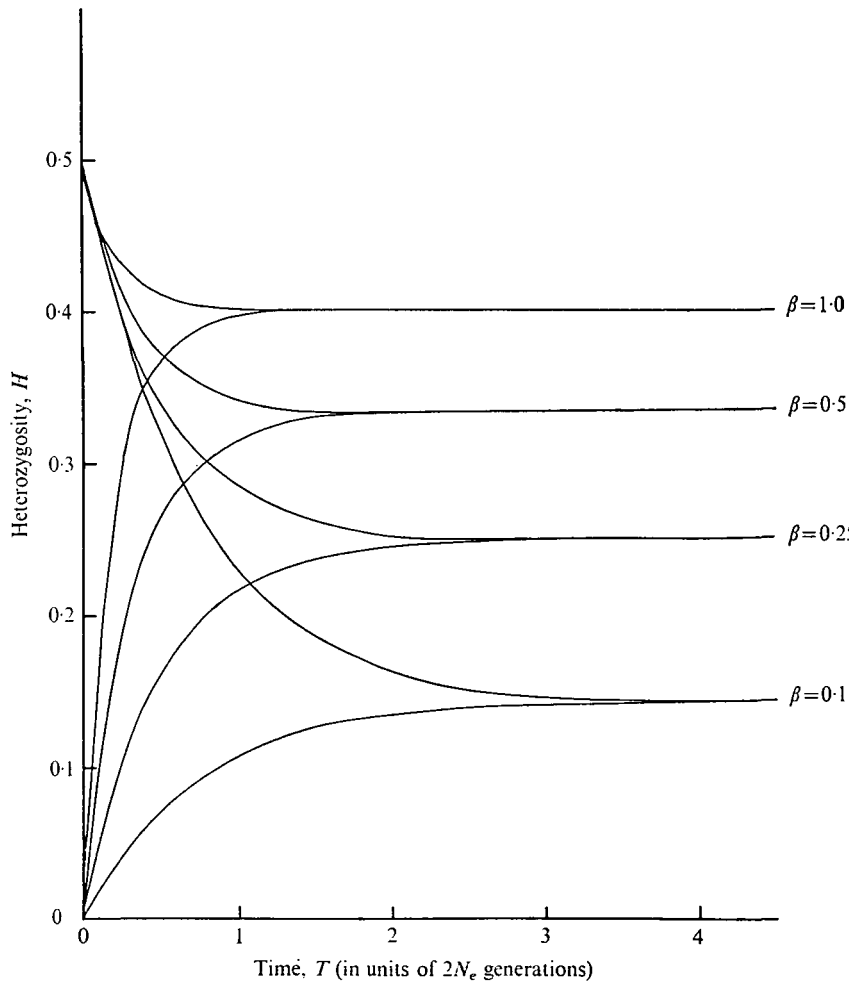

Fig. 4. The heterozygosity is plotted against time (where one unit is twice the colony size number of generations) in the two-allele, one-colony case for different values of the mutation factor, β. Initial conditions are either all of one allele (i.e. $H = 0$) or half of each allele (i.e. $H = 0{\cdot}5$).

https://doi.org/10.1017/S0016672300015548 Published online by Cambridge University Press

where $\boldsymbol{\gamma}_i$ and $\boldsymbol{\delta}_i$ are the ith column and row eigenvectors of **MA**, λ_i is the ith eigenvalue of **MA** and **E** is a column vector of arbitrary constants. For this particular **MA**, all the λ_i's are real and distinct. In fact

$$\lambda_i = -(1+2\beta) + 2\beta \cos\left(\frac{2\pi i}{K}\right) \quad (i = 0, \ldots, l),$$

and

$$\boldsymbol{\gamma}_i' = \left(1, \cos(2\pi i/K), \cos(4\pi i/K), \ldots, \cos\left(\frac{2l\pi i}{K}\right)\right).$$

These eigenvectors are not necessarily orthonormal. **E** can be found from the initial conditions and thus X_0 and H can be calculated.

The rate of convergence to steady state is determined by the λ_i's. The dominant eigenvalue is -1, i.e. is independent of K, the number of alleles. Thus, one obtains similar convergence curves for different K. Fig. 4 shows the results for $K = 2$, starting at (i) $H = 0$ (i.e. all of 1 allele) and (ii) $H = 0{\cdot}5$ (half of each allele).

7. CONCLUSION

One can see from the results how powerful mutation and migration factors are in influencing the expected steady state heterozygosity, even if they are relatively small. It is difficult to find data to prove or disprove the above theoretical solutions because of difficulty in the estimation of the parameters of the system. Also, the last section provides a warning, in that one must be careful only to try to apply this type of solution to populations where the parameters are constant over the relatively long time needed to reach steady state. Thus this model would be quite inappropriate for human populations.

That selection exists is an indisputable fact, but it is my opinion that most alleles are selectively neutral and that genetic variability is maintained by mechanisms similar to the above models. One would expect that in practice the heterozygosity would be slightly lower than the theoretical value, due to the presence of a few advantageous or deleterious alleles.

I would like to acknowledge the help and encouragement of Professor M. S. Bartlett and Dr M. G. Bulmer.

REFERENCES

KIMURA, M. & WEISS, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics* **49**, 561–576.

LATTER, B. D. H. (1973). The island model of population differentiation: a general solution. *Genetics* **73**, 147–157.

OHTA, T. & KIMURA, M. (1971). Linkage disequilibrium between 2 segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* **68**, 571–580.

OHTA, T. & KIMURA, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research* **22**, 201–4.

WEHRAHN, C. F. (1974). The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. (To be published in *Genetics*.)